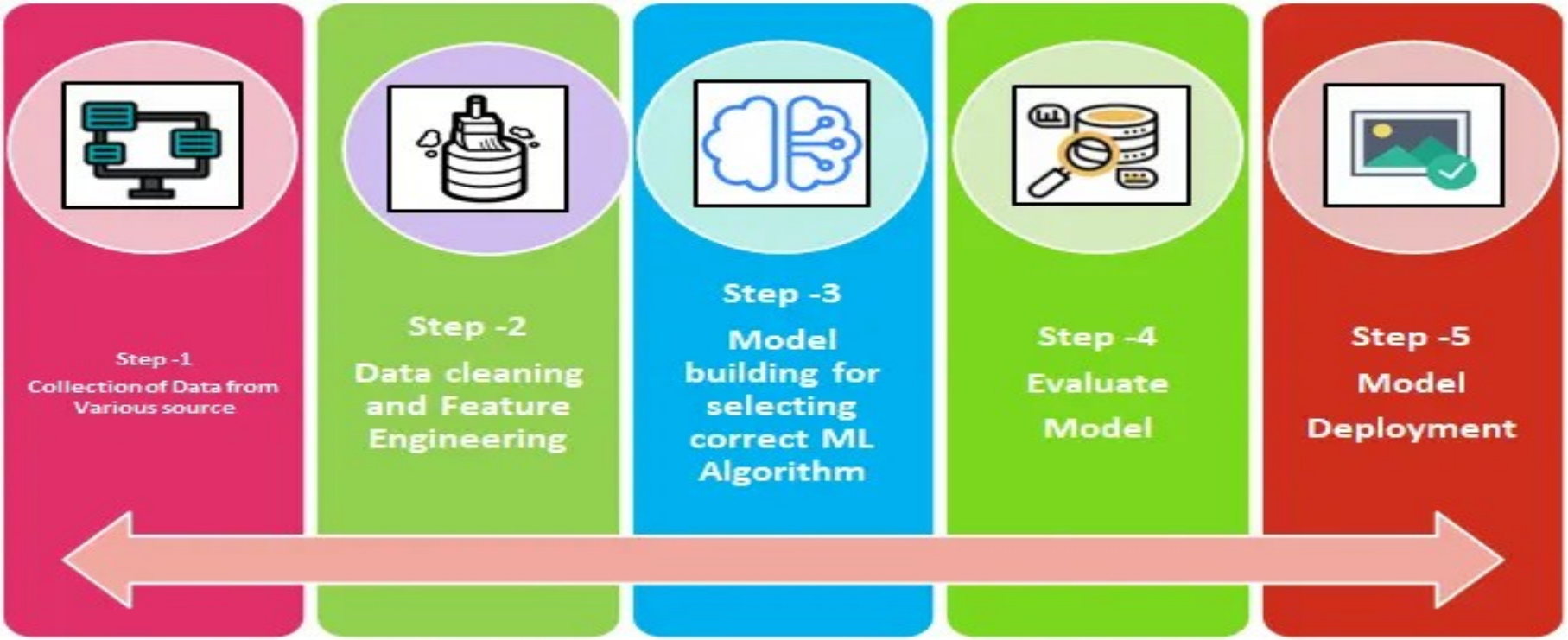
Chapter-4 Machine Learning- Supervised Learning

Basic Steps In ML

1. **Data collection**
   "training data", **mostly** with "labels" provided by a "teacher";
2. **Data preprocesing**
   Clean data to have homogenity
3. **Feature engineering**
   Select represenatative features to improve performance
4. **Modeling**
   choose the class of models that can describe the data
5. **Estimation/Selection**
   find the model that best explains the data: simple and fits well;
6. **Validation**
   evaluate the learned model and compare to solution found using other model classes;
7. **Operation**
   Apply learned model to new "test" data or real world instances

Basic Steps In ML



Common Terms

Features and Labels:

Features: These are the **input variables or characteristics** that the machine learning algorithm uses to make predictions.

Features provide the information on which the model's predictions are based.

The quality and relevance of features significantly **impact the performance of the machine learning model.**

House Price Prediction: Square footage, number of bedrooms, location,

number of bathrooms, presence of a garage.

Email Spam Classification: Email content, sender's address, presence of certain keywords.

Image Classification: Pixel values of an image, color distribution, texture features.

Common Terms

- **Labels**, also known as the **target variable or output variable**, represent the desired **outcome or prediction** that the model aims to achieve.

• Labels are the values that the model is **trying to predict**.

• The model's performance is assessed based on how well it predicts or approximates these labels.

House Price Prediction: Label: The actual price of the house.

Email Spam Classification: Label: Spam or not spam.

Image Classification: Label: Object categories (e.g. cat, dog, car).

Common Terms

- **Training Data:**

The training data is a subset of the available dataset that is used to train the machine learning model

- During training, the model adjusts its parameters based on this data to make accurate predictions.

Common Terms

- **Testing Data:**

- Once the model is trained on the training data, it is evaluated on a separate subset of data that was not used during the training process

- This testing data allows assessing how well the model generalizes to new, unseen data. Provides an unbiased evaluation of the model's ability to generalize.

- Helps identify if the model has **overfitting or underfitting** to the training data and whether it can make accurate predictions on real-world examples.

Common Terms

- **Overfitting**

- This phenomenon occurs when a model performs **really well on the data that we used to train** it but it **fails to generalise** well to new, unseen data. due to noise , and the model learned to predict specific inputs rather than the predictive parameters helps to make correct predictions

Under-fitting

- the model has **poor performance even on the data** that was used to **train it**. In most cases, underfitting occurs because the model is **not suitable for the problem** you are trying to solve

Data set Preparation

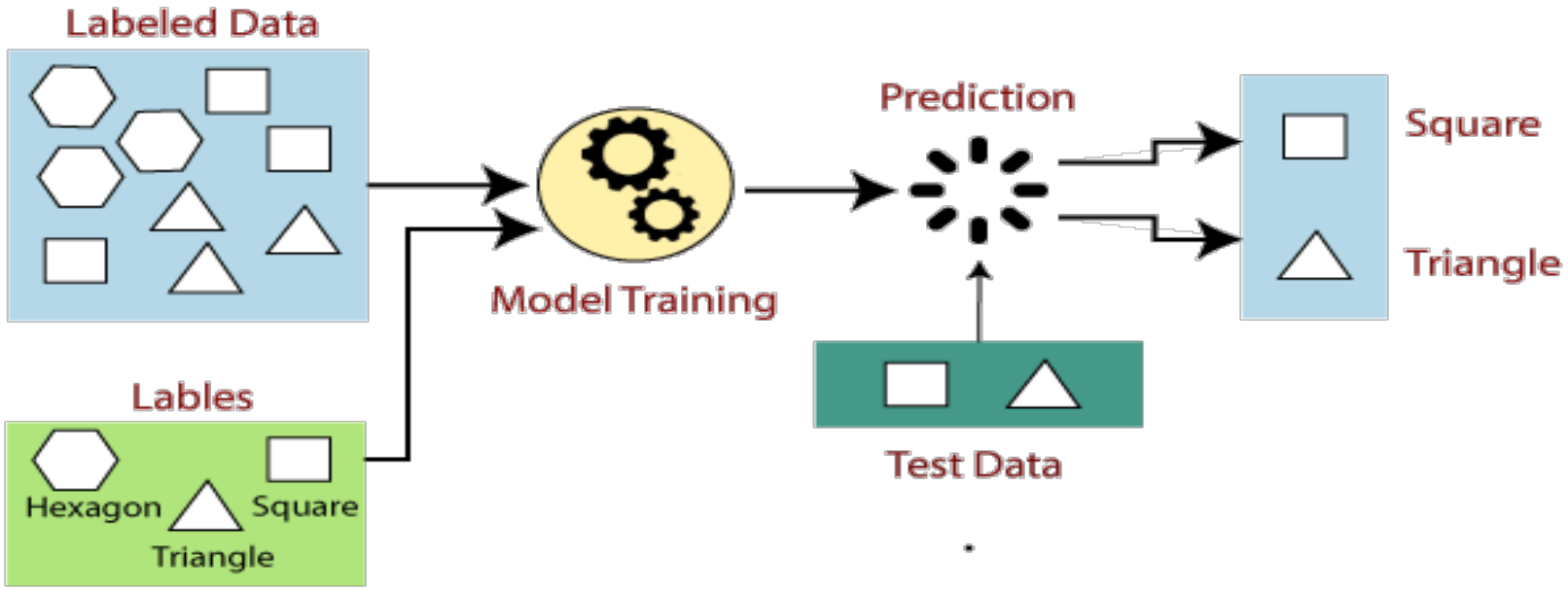
- To overcome over and under fitting try d/t approach of splitting
- simplest way to split the modelling dataset into **training and testing sets** is to assign **two thirds of the data** for training and rest for testing

Supervised ML

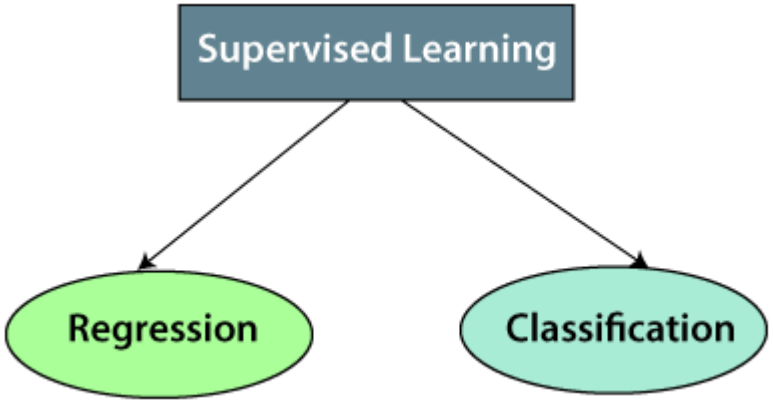
Supervised learning involves training an algorithm on a labeled dataset, where input data is paired with corresponding output labels.

- The goal is to learn a mapping from input to output based on provided labelled examples.

Supervised Learning



Supervised ML algorithms



Regression

Regression algorithms are used if there is a relationship between the input variable and the output variable.

- It is used for the prediction of **continuous variables**, such as Weather forecasting, Market Trends, etc. Below are some popular Regression algorithms which come under supervised learning: examples
  - Linear Regression
  - Regression Trees
  - Non-Linear Regression
  - Polynomial Regression

Linear Regression

- It is one of the very simple and easy algorithms which works on regression and shows the relationship between the continuous variables.

We should know that regression is a statistical method. It is used in finding relationships between variables.

- Linear regression is one of the regression-based algorithms in ML. It shows a linear relationship between its variables.

Assume some company x spent the following cost for advertisement and ge

the company wants to do the advertisement of $200 in the year 2019

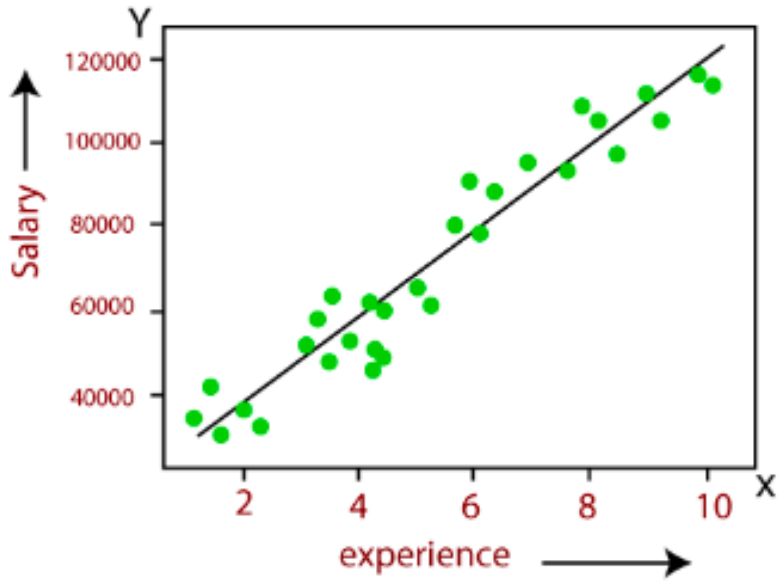
and wants to know the prediction about the sales for this year.

Advertisement	Sales
$90	$1000
$120	$1300
$150	$1800
$100	$1200
$130	$1380
$200	??

Example :

Linear reg cont...

2.Here we are predicting the **salary of an employee** on the basis of the year of experience.

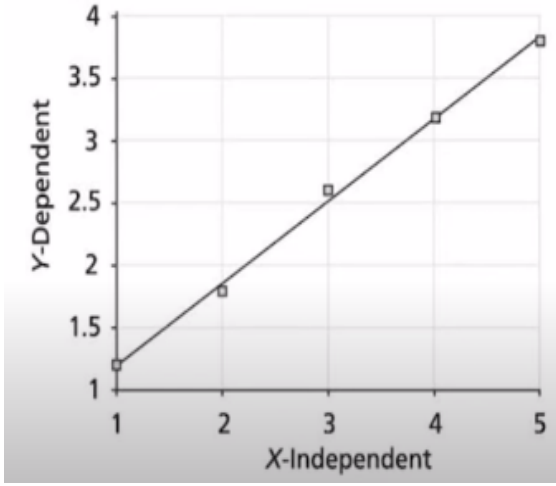


LR problem cont..

- Let us consider an example where the five weeks' sales data (in Thousands) is given as shown in Table.
- Apply linear regression technique to predict the 7^{th} and 12^{th} week sales.

x_i (Week)	y_j (Sales in Thousands)
1	1.2
2	1.8
3	2.6
4	3.2
5	3.8

Just finding the best fitting line



x_i (Week)	y_j (Sales in Thousands
1	1.2
2	1.8
3	2.6
4	3.2
5	3.8

Formula

$$y = \alpha + \beta x$$

β = slope

α = y-intercept

y = y- coordinate

x = x-coordinate

- Linear regression equation is given by
- $y = a_0 + a_1 * x + e$

- *where*
- $a_1 = \frac{(\overline{xy}) - (\bar{x})(\bar{y})}{\overline{x^2} - \bar{x}^2}$
- $a_0 = \bar{y} - a_1 * \bar{x}$

So/n cont...

- Here, there are 5 items, i.e., i = 1, 2, 3, 4, 5.

	x_i (Week)	y_j (Sales in Thousands)	x_i^2	$x_i * y_j$
	1	1.2	1	1.2
	2	1.8	4	3.6
	3	2.6	9	7.8
	4	3.2	16	12.8
	5	3.8	25	19
Sum	15	12.6	55	44.4
Average	$\bar{x} = 3$	$\bar{y} = 2.52$	$\overline{x^2} = 11$	$\overline{xy} = 8.88$

- *where*
- $a_1 = \frac{(\overline{xy}) - (\bar{x})(\bar{y})}{\overline{x^2} - \bar{x}^2}$
- $a_0 = \bar{y} - a_1 * \bar{x}$

Get correct regression line

- $\bar{x} = 3$ $\bar{y} = 2.52$ $\overline{x^2} = 11$ $\overline{xy} = 8.88$

- $a_1 = \frac{(\overline{xy})-(\bar{x})(\bar{y})}{\overline{x^2}-\bar{x}^2} = \frac{8.88-3*2.52}{11-3^2} = 0.66$
- $a_0 = \bar{y} - a_1 * \bar{x} = 2.52 - 0.66 * 3 = 0.54$
- **Regression equation is**
- $y = a_0 + a_1 * x$
- $y = 0.54 + 0.66 * x$

Linear Regression

- **Regression equation is**
- $y = a_0 + a_1 * x$
- $y = 0.54 + 0.66 * x$
- The predicted 7th week sale (when x = 7) is,
- **y = 0.54 + 0.66 x 7 = 5.16**
- the predicted 12th week sale (when x = 12) is,

Practical ML-Prediction problem

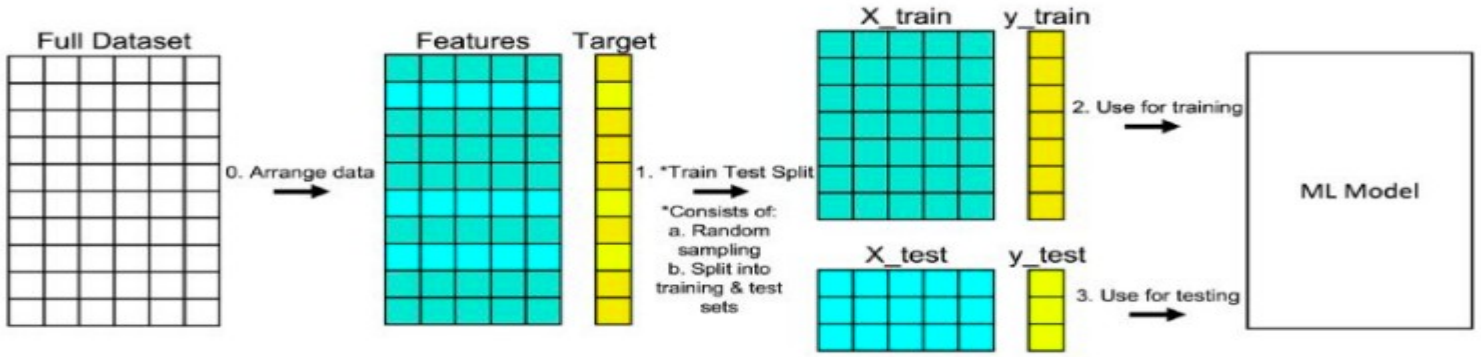
Step-1: Data Loading

Step 2:Identify Independent /Dependent variables */predictions*

Step 3: Split the data to train/test the ML algorithms

Step-4:train the model and test it

Train test split is a model validation procedure that allows you to simulate how a model would perform on new/unseen data. Here is how the procedure works:



Change this to practical ML

```
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0, train_size =
.75)
```

The random_state:

is a pseudo-random number parameter that allows you to **reproduce the same train test split** each time you run the code.

Select data set randomly before splitting just put /shuffling +ve integer commonly 42 , 2 ,0 default None

=unless you put random state you will get d/t values

Exercise : please split the following data with random state

X=[10,20,30,40,50,60,80,90,100]

Y=[1,0,1,4,5,6,7,8,9,10]

Sample Splitting task

```
from sklearn.model_selection import train_test_split
x=[10,20,30,40,50,60,80,90,100,200]
y=[1,0,1,4,5,6,7,8,9,10]
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=2)
print("x_train",x_train)
print("x_test",x_test)
print("y_train",y_train)
print("y_test",y_test)
```